

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2003-76592
(P2003-76592A)

(43) 公開日 平成15年3月14日 (2003.3.14)

(51) Int.Cl.	識別記号	F I	テマコード* (参考)
G 0 6 F 12/00	5 3 5	G 0 6 F 12/00	5 3 5 Z 5 B 0 6 5
	5 3 1		5 3 1 D 5 B 0 8 2
	5 3 3		5 3 3 J
3/06	3 0 4	3/06	3 0 4 B

審査請求 未請求 請求項の数 6 O L (全 15 頁)

(21) 出願番号 特願2001-266629 (P2001-266629)

(22) 出願日 平成13年9月4日 (2001.9.4)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 日野 直樹

神奈川県小田原市中里322番地2号 株式会社日立製作所 SANソリューション事業部内

(72) 発明者 占部 喜一郎

神奈川県小田原市中里322番地2号 株式会社日立製作所 RAIDシステム事業部内

(74) 代理人 100075096

弁理士 作田 康夫

最終頁に続く

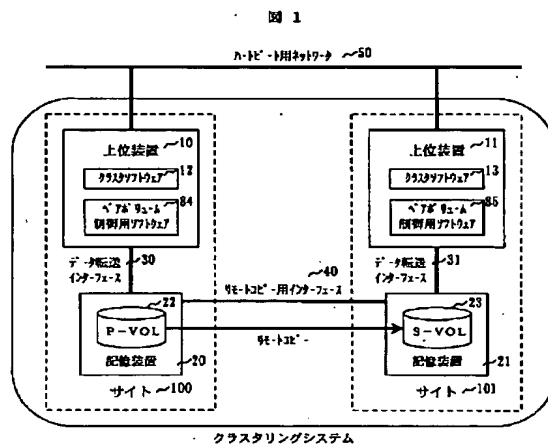
(54) 【発明の名称】 データ格納システム

(57) 【要約】

【課題】リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用するデータ格納システムを実現する。

【解決手段】分散して設置された2台のリモートコピー機能を持つ記憶装置20、21間で作成されたペアボリューム22、23を、2台以上の上位装置10、11から共用する構成において、上位装置10、11上のクラスタソフトウェア12、13から発行されるリザーブコマンド等のSCSIコマンドを一旦、ペアボリューム制御用ソフトウェア84、85が受け、当該ペアボリュームが唯一で共通の「状態」となるようにして操作するペアボリューム制御コマンドとを、リザーブコマンド等のSCSIコマンドと組み合わせて発行することにより、ペアボリュームを排他的使用する上位装置を一意に決定する。

【効果】リモートコピー機能によって生成されたペアボリュームを、クラスタリングシステムの判定用ボリュームとして使用することが可能となる。



【特許請求の範囲】

【請求項1】分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、2台以上の上位装置から共用するデータ格納システムにおいて、1台の前記上位装置が前記ペアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否する制御用ペアボリュームと、データ格納用ペアボリュームから成り、前記制御用ペアボリュームに2台以上の前記上位装置から更新要求があった場合に、前記制御用ペアボリュームに最も早くアクセスを行った前記上位装置が前記制御用ペアボリュームを排他的に使用する権利を獲得するデータ格納システム。

【請求項2】請求項1記載のデータ格納システムにおいて、ペアボリュームの属性及び、前記ペアボリュームの属性を遷移させるペアボリューム制御コマンドを発行する、ペアボリューム制御用ソフトウェアを使用することで、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項3】請求項1または請求項2のデータ格納システムにおいて、上位装置上の汎用アプリケーションと連携して動作し、前記記憶装置に対し前記ペアボリューム制御コマンドを発行する前記ペアボリューム制御用ソフトウェアを使用する事で、前記制御用ペアボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【請求項4】2台以上の上位装置と、2台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化する、データ格納システム。

【請求項5】2台以上の上位装置と、2台の記憶装置がデータ転送インターフェース及びネットワークインターフェースで接続されるクラスタリングシステムにおいて、前記上位装置もしくはネットワークインターフェースの障害発生時に前記上位装置の専用使用権を一意に決定するための前記記憶装置内の判定用ボリュームをリモートコピー機能で二重化し、複数の前記上位装置から前記判定用ペアボリュームを1つのボリュームとして認識するクラスタリングシステムを実現するデータ格納システム。

【請求項6】分散して設置されたデータを格納する第一の記憶装置と第二の記憶装置間でリモートコピー機能により作成されたペアボリュームを、2台以上の上位装置から共用するデータ格納システムにおいて、ペアボリュームの専有使用を目的として前記上位装置から発行されたSCSIリザーブコマンドを前記ペアボリューム制御

用ソフトウェアを介して、前記記憶装置内のボリュームに発行し、前記SCSIリザーブコマンドに対する前記ボリュームから正常終了のレスポンスを前記ペアボリューム制御用ソフトウェアがうけて、ペアボリューム制御コマンドを前記ボリュームに発行し、前記ボリュームの属性の確認及び、前記属性に応じて、前記ボリュームの前記属性を遷移させる前記コマンドを発行し、正常終了した場合のみ、前記SCSIリザーブコマンドに対する正常終了を前記上位装置に応答し、前記ボリュームに対して更新要求を発行できる上位装置を排他的に決定するデータ格納システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は分散して設置された2台以上の上位装置と2台の記憶装置から構成されるデータ格納技術に関する。

【0002】更に具体的には分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成された2重化ボリュームの排他的制御を行うデータ格納技術に関する。

【0003】

【従来の技術】システムを止めることなく24時間サービスを継続させるためのさまざまなIT技術がいま、特にEビジネス化を推進する企業の関心を集めている。その中で高可用性を実現する技術の一つがクラスタリングである。複数台のシステムを組み合わせ、一部システムの障害時にもシステム全体が停止せず稼働できるように構築することで、データベース業務を行うシステム等で主に使われている。例えば、小規模システムの例では数台のサーバーで1つの共有ディスクを利用する構成をとり、サーバーと共有ディスク間はSCSIインターフェースを用い、この構成で数台のサーバーが扱っているデータを共有ディスクに置いておけば、一台のサーバーがダウンしても、他の一台のサーバーが共有ディスクのデータを引き継ぎ、処理を継続することができる。これをフェイルオーバーという。このようにフェイルオーバーを行うクラスタリングシステムでは、クライアントからはただ一つのサーバーを引き続き利用しているように業務を行える。

【0004】また、Microsoft社のクラスタリング製品であるMSCS（マイクロソフト・クラスタ・サーバ）ではクラスタリングシステム内でフェイルオーバーを行う上位装置を一意に決定する為に、クラスタ構成情報を管理するクォラムリソースと呼ばれる一つの判定用ボリュームを持つ。この判定用ボリュームに対して排他制御を行うことで、そのデータの一貫性を維持できる。

【0005】すなわち、複数の異なるプロセスが同時に一つのリソースに対して重複処理をしてしまうことにより、データの一貫性が損なわれるという問題を回避している。

【0006】次に、2台以上の上位装置と、2台の記憶装置がSCSIインターフェースによって接続される場合のクラスタリングシステムを例として説明する。クラスタリングシステムでは、各々の上位装置間をネットワークインターフェースで接続し、各上位装置は互いにメッセージを発信してハートビート通信を行う事で、お互いの移動状況を監視する。

【0007】さらに、クラスタリングシステムは上位装置の障害又はネットワークインターフェースの障害が発生し、ハートビート通信が不可能になったことを検出して、個々の上位装置が他の上位装置のリソース、アプリケーション及びサービスの引き継ぎを行うかどうかの判断を行う。その判断を行う唯一の決定要素としてクラスタ構成情報を管理する1つの判定用ボリュームを記憶装置内に設ける。

【0008】一般にSCSIインターフェースを使用するシステムにおいては、複数のホストのうちひとつのホストがひとつのターゲット（例えば、磁気ディスクドライブ）を排他制御（専用使用）をする場合、最初にSCSIリザーブコマンドを発行することにより、そのターゲットをリザーブでき専用使用が可能となる。

【0009】上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置はクラスタリングシステム内での専用使用権を獲得する為に、判定用ボリュームに対してSCSIリザーブコマンドを発行し、専用使用権を手に入れることを試みる。判定用ボリュームへのSCSIリザーブコマンドの実行が成功した上位装置は他の上位装置のアプリケーション、サービス及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバーを実行する。一方、判定用ボリュームに対するSCSIリザーブコマンド実行が失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する。

【0010】このように、クラスタリングシステムは、一台の上位装置がクラスタ構成情報を管理する判定用ボリュームを専用使用することで実現されている。

【0011】

【発明が解決しようとする課題】特に大量のデータをリアルタイムで処理する必要がある、大規模なオンライントランザクションシステムにおいてはシステムを止めないために、サーバーをはじめディスク装置や電源などを多重化し、一つが故障しても他方が処理を引き継ぐ、フェイルオーバー機能をサポートしたクラスタリングシステムが必要になる。

【0012】従来技術によれば、クラスタリングシステムにおける判定用ボリュームは全ての上位装置がアクセスすることができる唯一、1つのディスクボリュームによって構成されている。

【0013】しかし、複数の上位装置が分散配置されるクラスタリングシステムにおいては、1つの判定用ボリ

リュームを全ての上位装置がアクセスできるように設置するには、上位装置と判定用ボリュームを有する記憶装置間のインターフェース、例えばSCSIインターフェース等、接続用のケーブルの接続距離の制限がありシステム構築が出来ない場合があるという問題があった。

【0014】これを解決する手段として、災害バックアップシステムや遠隔地へのデータ送信システムを構築できるリモートコピー機能をクラスタリングシステムに用いることが考えられる。

【0015】リモートコピー機能では上位装置を経由せずに（ホストフリー）、記憶装置間のリモートコピーを実現する。リモート側への更新データ反映順序を守るため、リモート側のコピーデータの論理的・一貫性の保証が可能である。ここで、リモートコピー機能により分散配置された記憶装置間で二重化されたペアボリューム（それぞれを正ボリューム及び副ボリュームと呼び、正ボリュームを有する記憶装置及び副ボリュームを有する記憶装置がある）を作成し、ペアボリュームの正ボリュームもしくは副ボリュームは同一のデータを持つことが保証されているので、ある一台の上位装置のみが正ボリュームと副ボリュームのどちらかを判定用ボリュームとして専用使用する上位装置を決定する方法が考えられる。

【0016】しかし、分散配置された各システムはそれぞれ、そのシステム内でそれぞれのSCSIインターフェースを使用しているシステムゆえ、各上位装置はそれぞれのシステム内でSCSIリザーブコマンドを発行し、実行が成功した上位装置がボリュームを排他制御できることになり、正ボリュームを専用使用する上位装置及び副ボリュームを専用使用する2台の上位装置が存在することになる。

【0017】すなわち、正ボリュームもしくは副ボリュームを排他的に使用する目的でSCSIリザーブコマンドを発行しただけでは、正ボリュームもしくは副ボリュームは別々にリザーブすることができる為に、判定用ボリュームとして排他使用ができないという問題があった。

【0018】

【課題を解決するための手段】前述に示す通り、リモートコピー機能を用いて判定用ボリュームを設定する場合、ペアを構成する正ボリューム及び副ボリュームは各上位装置から別々のボリュームとして認識され、各上位装置が正ボリューム及び副ボリュームに対し、SCSIリザーブコマンドを発行した場合、正ボリューム及び副ボリュームを個々にリザーブすることができ、結果として複数の上位装置が同じアプリケーション及びサービスを同時に移動してしまい、クラスタリングシステムとしての機能をはたさないことになる。

【0019】従って、リモートコピー機能をもつ2台の記憶装置間で作成されたペアボリュームを、複数の上位装置からなるクラスタリングシステムにおいて1つの判

定用ボリュームとして認識させ、クラスタリングシステム内の複数の上位装置間で当該ベアボリュームに対し更新要求を発行できる上位装置を排他的に決定する手法を実現する為に、本発明では上位装置上のクラスタソフトウェアから発行されるSCSIコマンドのリザーブコマンドを、ベアボリュームの「状態」(属性)を操作するベアボリューム制御用ソフトウェアが一旦受け、ベアボリューム制御用ソフトウェアが発行するベアボリューム制御コマンドと組み合わせて記憶装置に発行し、そのコマンドに対する記憶装置のレスポンスが、最初にリザーブコマンドを発行した上位装置のみ正常終了とすることで、ベアボリュームを排他的使用する上位装置を一意に決定することが可能となるデータ格納システムを提案する。

【0020】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて説明する。尚、クラスタリングシステムを利用した実施例としたが、本発明の適用範囲はクラスタリングシステムに限ったものではない。また、2台の上位装置からなるクラスタリングシステムを実施例として用いたが、上位装置は3台以上あっても構わない。

【0021】図1は本発明の一実施形態である。図1はそれぞれ上位装置と記憶装置からなる2つのサイト100、101をネットワークインターフェースによって接続し、クラスタリングシステムを実現する構成図である。10、11はCPUを持つ上位装置であり、上位装置10、11上ではクラスタソフトウェア12、13及び本発明で提案するベアボリューム制御用ソフトウェア84、85等が動作する。また、上位装置10、11はハートビート用ネットワーク50によって接続され、定期的にお互いの移動状況を監視する。

【0022】20、21は上位装置10、11からのデータを格納しておく記憶装置であり、データ転送インターフェース30、31を介して接続される。また、記憶装置20、21は複数のボリュームから構成され、上位装置10、11上にインストールされたアプリケーションからのデータ更新及び参照要求を受けることができる。

【0023】さらに、記憶装置20、21はリモートコピー機能を持ち、一方の記憶装置上の正ボリューム(P-VOL)22のミラーをもう一方の記憶装置上の副ボリューム(S-VOL)23に構築することができる。上位装置10で動作するアプリケーションは記憶装置20内の正ボリューム(P-VOL)22に対してデータの更新要求を行い、その更新内容がリモートコピー用インターフェース40を通じて自動的にもう一方の記憶装置内の副ボリューム(S-VOL)23に反映される。

【0024】図2はクラスタリングシステムの内部構造を示す。上位装置10、11はそれぞれクラスタソフトウェア80、81、アプリケーション82、83及びベ

アボリューム制御用ソフトウェア84、85から構成される。ベアボリューム制御用ソフトウェア84、85はクラスタソフトウェア80、81と標準デバイスドライバ90、91の中間に位置する。

【0025】クラスタソフトウェア80、81はベアボリューム制御用ソフトウェア84、85が独自に定義したデバイスファイルに対してSCSIコマンドを発行し、それを受け取ったベアボリューム制御用ソフトウェア84、85は、必要に応じてベアボリューム制御コマンドと組み合わせたコマンドを標準デバイスドライバ90、91経由でクラスタ構成情報を管理する判定用ベアボリューム60、61へ発行する。

【0026】また、記憶装置20、21はクラスタの専有使用権を確保する為に構築された判定用ベアボリューム60、61、及びアプリケーション用のデータを格納するデータボリューム70-73で構成される。リモートコピー機能によって各ボリュームは二重化されており、リモートコピー機能の仕様により、各アプリケーションは正ボリューム(P-VOL)のみに更新要求を発行する事ができる。

【0027】クラスタソフトウェア80と81はネットワークを介してハートビート通信を行うことで定期的に相手側上位装置の移動状況を監視する。上位装置10又は11の障害、又はハートビートネットワーク50(図1)の障害によりハートビート通信が不可能となった場合、クラスタソフトウェア80、81はベアボリューム制御用ソフトウェア84、85を介して、記憶装置20、21内に構築された判定用ベアボリューム60、61にそれぞれSCSIリザーブコマンドを発行する。

【0028】そして、判定用ベアボリュームをリザーブすることに成功した上位装置はクラスタソフトウェアによりクラスタリングシステム内に登録された全リソース及びアプリケーションの引き継ぎを行う。判定用ベアボリュームをリザーブすることに失敗した他の上位装置はクラスタソフトウェアによりリソースを解放し、アプリケーションを停止する。

【0029】ここで、図3に示すベアボリュームの状態遷移について詳細説明する。リモートコピー機能によって生成された各ベアボリュームは一般的に「非ベア状態」、「コピー状態」、「ベア状態」、「サスペンド状態(正常)」、「サスペンド状態(異常)」の状態を持ち、上位装置上にインストールされたベアボリューム制御用ソフトウェアがベアボリューム制御コマンドを記憶装置へ発行することで、リモートコピー用インターフェース40を通じて自動的に各々のベアボリュームの状態を遷移させることができる。

【0030】「非ベア状態」は2つのボリューム間にベアが形成されていない状態、「コピー状態」は2つのボリューム間で正ボリュームから副ボリュームへコピー実行中でベアボリュームのデータはまだ全て同等ではない状

態。コピー完了時、ボリュームは「ベア状態」に変わる。「ベア状態」では、上位装置から正ボリュームへの全ての更新は副ボリュームに対して二重化される。「サスペンド状態(正常)」は正ボリュームに対する更新を副ボリュームに反映しない状態。例えばベアボリューム間で正ボリュームのみ更新をかけたい時、正、副ボリュームを本状態に設定する。

【0031】「サスペンド状態(異常)」は前述と違い、何らかの障害検知で、ベアボリューム間で内容の更新を保持できない時、正、副ボリュームを本状態に設定する。各ベアボリュームの状態を遷移させるベアボリューム制御コマンドは正ボリュームまたは副ボリュームのどちらに接続された上位装置からでも実行可能である。

【0032】例えば、ベアボリュームが「サスペンド状態(正常)」であった場合、正ボリューム側の上位装置からベア再同期コマンドを発行し「ベア状態」に遷移させる事も可能であるし、副ボリューム側の上位装置からベア再同期コマンドを発行し、「ベア状態」に遷移させることもできる。さらに、副ボリューム側の上位装置からベアスワップ再同期コマンドを発行し、副ボリュームを正ボリュームにスワップした上で「ベア状態」に遷移させることもできる。

【0033】また、ベアボリューム制御コマンドの実行結果は実行先のベアボリュームの状態に依存するという性質を持つ。以下詳細説明する。

【0034】図4にベアボリューム制御コマンドであるベア再同期コマンド及びベアスワップ再同期コマンド発行時の記憶装置からのレスポンスの例を示す。例えば、実行コマンドとしてベア再同期コマンドを発行した場合、実行先のボリュームが正ボリュームでかつ「サスペンド状態(正常)」でなければ記憶装置からのレスポンスは正常終了とはならない。

【0035】もし、実行先のボリュームが「コピー状態」もしくは「ベア状態」であった場合はそのベア再同期コマンドは異常終了する。また、ベアスワップ再同期コマンドは実行先のベアボリュームが副ボリュームであり、かつ「サスペンド状態(正常)」でなければ記憶装置からのレスポンスは正常終了とはならない。

【0036】もし、実行先のボリュームが「コピー状態」もしくは「ベア状態」であった場合はそのコマンドは異常終了する。このベアボリュームの性質を用いて、ベアボリューム制御用ソフトウェアはクラスタソフトウェアから発行されたSCSIコマンドを一旦受け取り、本SCSIコマンドを分析した結果、リザーブコマンド等のベアボリューム間で排他的な処理を必要とするSCSIコマンドであった場合はベアボリューム制御コマンド例えば、ベア再同期コマンドと組み合わせて記憶装置側へ発行する。

【0037】そして、ベア再同期コマンドが記憶装置内で実行され、正常にベアボリューム状態を遷移できた場

合は、正常終了ステータスをベアボリューム制御用ソフトウェアに送信する。さらに、正常終了ステータスを受け取ったベアボリューム制御ソフトウェアはクラスタソフトウェアに対して正常終了ステータスを返す。

【0038】正常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内で引き続き生存することが許可され、フェイルオーバーを実行する。一方、1つの上位装置から最初に発行されたベア再同期コマンドが記憶装置上で実行され、ベアボリュームの状態が「ベア状態」に遷移したことによって、他の上位装置から同様のベアボリューム制御コマンドが発行されたとしても図4に示すよう、「ベア状態」に発行されたベアボリューム制御コマンドは全て異常終了となる。

【0039】ベアボリューム制御コマンドの異常終了報告を受けたベアボリューム制御用ソフトウェアはSCSIコマンド要求が失敗したことの報告として、クラスタソフトウェアへ異常終了ステータスを返す。異常終了ステータスを受け取ったクラスタソフトウェアはクラスタリングシステム内での生存が許されず、担当しているアプリケーション及びサービスを全て停止させる。

【0040】図5、6で、本発明で提案するベアボリューム制御用ソフトウェアの動作についてSCSIリザーブコマンドを受けた場合とそれ以外のコマンドをうけた場合について更に詳細に説明する。ベアボリューム制御用ソフトウェア84はクラスタソフトウェア80から発行されたSCSIコマンドがリザーブコマンドであった場合(110)、判定用ベアボリュームの正ボリューム(P-VOL)または副ボリューム(S-VOL)に対し、SCSIリザーブコマンドを発行する(111)。

【0041】SCSIリザーブコマンドに対するボリュームからのレスポンスが正常終了であった場合(112)はボリュームが判定用ベアボリュームの正ボリューム(P-VOL)か副ボリューム(S-VOL)かを調査する為にボリューム制御用ソフトウェア84は記憶装置に対してボリューム属性確認コマンドを発行する(113)。判定用ボリュームが正ボリューム(P-VOL)であった場合は更にベア再同期コマンドを発行する(114)。

【0042】一方、判定用ボリュームが副ボリューム(S-VOL)であった場合はベアスワップ再同期コマンドを発行する(115)。ここで、ベア再同期コマンドとはサスペンド状態のベアボリュームを再同期しベア状態に遷移させるコマンドである。

【0043】一方、ベアスワップ再同期コマンドとはサスペンド状態のベアボリュームを再同期しベア状態に遷移させ、かつ正ボリュームを副ボリュームに、副ボリュームを正ボリュームにスワップするコマンドである。これにより、副ボリュームであった判定用ボリュームが正ボリュームに遷移する。

【0044】図4に示すように、ベア再同期コマンドは

コマンド発行配下のボリュームすなわち自ボリュームが正ボリューム (P-VOL) かつサスペンド状態 (正常) の場合のみ成功する (130)。

【0045】一方、ベアスワップ再同期コマンドは自ボリュームが副ボリューム (S-VOL) かつサスペンド状態 (正常) の場合のみ成功する (131)。

【0046】また、図3に示すように、ベア再同期コマンドまたはベアスワップ再同期コマンドのどちらかが発行され、実行状態になった場合に、ベアボリュームの状態は「サスペンド状態 (正常)」(120) から「コピー状態」(121) さらに、コピーが完了すると「ベア状態」(122) に遷移する。よって、図4に示すように、ベアボリュームの状態が「ベア状態」になった後で発行されたベア再同期コマンドまたはベアスワップ再同期コマンドは異常終了することになる。

【0047】このようにベアボリューム制御用ソフトウェアはベア再同期またはベアスワップ再同期コマンドが正常終了した場合に、SCSIリザーブコマンドが成功したことを上位装置へGoodステータス (成功) を返して報告する (117)。すなわちクラスタの専用使用権が取れたことを報告する。

【0048】一方、ベアボリューム制御コマンドが異常終了した場合は、リザーブに失敗した、すなわちクラスタの専用使用権が取れなかったことを報告する為に、上位装置に対して、Reservation Conflictステータス (失敗) を返す (116)。

【0049】判定用ボリュームへのリザーブ要求に対して、Goodステータスを得た上位装置はフェイルオーバを実行し、クラスタ内の全てのリソース、アプリケーション及びサービスを引き継ぐ。

【0050】一方、Reservation Conflictステータスを得た上位装置は自身が管理していた全てのリソースを解放し、かつ自身が管理していた全てのアプリケーション及びサービスを停止する。

【0051】また図6に示すように、クラスタソフトウェア80、81から発行されたSCSIコマンドがリザーブコマンド以外であった場合はベアボリューム制御用ソフトウェアはベアボリューム制御コマンドを付加せず、そのままの状態判定用ボリュームに当該SCSIコマンドを発行する。

【0052】図7から図13にて本発明をクラスタリングシステムに適用した場合の、システムが正常な状態から障害によりクラスタリングが機能する一連の動作について説明する。まず、最初の状態ではクラスタリングシステムは停止し、判定用ベアボリュームは「サスペンド状態 (正常)」を維持している (図7)。

【0053】ここで、ユーザがクラスタリングシステムを起動した場合、各上位装置内のクラスタソフトウェアはクラスタリングシステム内の専用使用権を獲得する為に、ベアボリューム制御用ソフトウェアに対してSCS

Iリザーブコマンドを発行する。SCSIリザーブコマンドを受領したベアボリューム制御用ソフトウェアは、判定用ベアボリュームの正ボリューム (判定用 (P)) または副ボリューム (判定用 (S)) に対し、SCSIリザーブコマンドを発行し、専用使用権争いを行う。本例では各システム内に一台の上位装置のみ存在するので、上位装置10、11が発行するSCSIリザーブコマンドに対する各ボリュームからベアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。

【0054】次に、判定用ベアボリュームの正ボリューム (判定用 (P)) 側に接続するベアボリューム制御用ソフトウェアはベア再同期コマンドを発行する。一方、判定用ベアボリュームの副ボリューム (判定用 (S)) 側に接続するベアボリューム制御用ソフトウェアはベアスワップ再同期コマンドを発行する。

【0055】本例では上位装置10のベア再同期コマンドが上位装置11のベアスワップ再同期コマンドより先に発行され、クラスタの専用使用権を獲得したとする。専用使用権を獲得した上位装置10内のクラスタソフトウェアはWriteコマンドを判定用ベアボリュームに対して発行し、クラスタ管理情報の初期化を行う (図8)。

【0056】次に、各上位装置内のクラスタソフトウェアはユーザにて予め設定されたデータボリュームをマウントし、そのデータボリュームに対するアプリケーションを起動する (図9)。

【0057】さらに、判定用ベアボリュームに対してはシステムの障害に備え、ベアボリューム制御用ソフトウェアを用いて判定用ベアボリュームを分割することで「サスペンド状態 (正常)」に遷移させ (図10)、システムは通常運用に入る (図11)。

【0058】通常運用中に上位装置の障害又はネットワークの障害等が発生し、ハートビート通信が不可能になった場合、各々の上位装置内のクラスタソフトウェアはクラスタリングシステム内での専用使用権を獲得する為に、SCSIバスデバイスリセットを発行し、データ転送インターフェース上をリセットした後、ベアボリューム制御用ソフトウェアに対してSCSIリザーブコマンドを発行する。SCSIリザーブコマンドを受領したベアボリューム制御用ソフトウェアは、判定用ベアボリュームの正ボリューム (判定用 (P)) または副ボリューム (判定用 (S)) に対し、SCSIリザーブコマンドを発行し、専用使用権争いを行う。

【0059】本例では各サイト内に一台の上位装置のみ存在するので、上位装置10、11が発行するSCSIリザーブコマンドに対し、各ボリュームからベアボリューム制御用ソフトウェアに対するレスポンスはそれぞれ正常終了となる。次に、判定用ベアボリュームの正ボリューム (判定用 (P)) 側に接続するベアボリューム制御用ソフトウェアはベア再同期コマンドを発行する。

【0060】一方、判定用ベアボリュームの副ボリューム(判定用(S))側に接続するベアボリューム制御用ソフトウェアはベアスワップ再同期コマンドを発行する。本例では上位装置10のベア再同期コマンドが上位装置11より先に発行され、クラスタの専用使用権を獲得したとする(図12)。

【0061】専用使用権を獲得した上位装置10内のクラスタソフトウェアは他の上位装置のアプリケーション及びディスクボリューム等のリソースを全て引き継ぎ、フェイルオーバーを実行する。一方、判定用ボリュームに対するSCSIリザーブコマンドが失敗した上位装置はその上位装置上で実行していたアプリケーション及びサービスを全て停止する(図13)。

【0062】このように、本発明により分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成されたベアボリュームを、1台の上位装置が当該ベアボリュームを排他的に占有し、他の上位装置からの更新要求を拒否することで、ベアボリュームを当該複数の上位装置に対して1つのボリュームとして認識させることができ、当該ベアボリュームをクラスタリングシステムの判定用ボリュームに適用することができる。

【0063】

【発明の効果】本発明は分散して設置された2台のリモートコピー機能を持つ記憶装置間で作成されたベアボリュームを、2台以上の上位装置から共用する構成において、1台の前記上位装置が前記ベアボリュームを排他的に占有し、他の前記上位装置からの更新要求を拒否することで、前記ベアボリュームを前記複数の上位装置に対して1つのボリュームとして認識させるデータ格納システムを実現する。

【図面の簡単な説明】

【図1】本発明の一実施形態を示すハードウェア構成図である。

【図2】リモートコピー機能のベアボリュームを判定用ボリュームに使用したクラスタリングシステムの構成図である。

【図3】ベアボリュームの状態遷移図である。

【図4】ベア再同期コマンド又はベアスワップ再同期コマンド発行する場合の、各ボリューム属性に対する記憶装置からのレスポンスを示す

【図5】SCSIリザーブコマンド受領時のベアボリューム制御用ソフトウェアの動作である。

【図6】SCSIリザーブコマンド以外受領時のベアボリューム制御用ソフトウェアの動作である。

【図7】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【図8】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【図9】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【図10】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【図11】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【図12】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

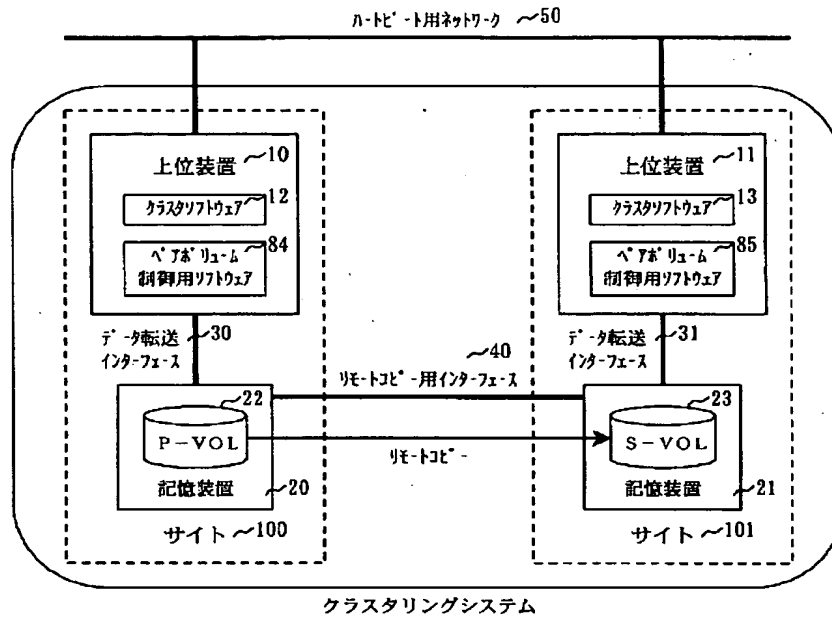
【図13】判定用ベアボリュームを持つクラスタリングシステムの動作を示す。

【符号の説明】

- 10、11 上位装置
- 12、13 クラスタソフトウェア
- 20、21 記憶装置
- 22 正ボリューム(P-VOL)
- 23 副ボリューム(S-VOL)
- 30、31 データ転送用インターフェース
- 40 リモートコピー用インターフェース
- 50 ハートビート用ネットワーク
- 60 判定用ベアボリュームの正ボリューム(P-VOL)
- 61 判定用ボリュームの副ボリューム(S-VOL)
- 70、73 データボリュームの正ボリューム(P-VOL)
- 71、72 データボリュームの副ボリューム(S-VOL)
- 80、81 クラスタソフトウェア
- 82、83 アプリケーション
- 84、85 ベアボリューム制御用ソフトウェア
- 90、91 標準デバイスドライバ
- 100、101 サイト
- 110 クラスタソフトウェアのSCSIリザーブコマンド発行
- 111 ベアボリューム制御用ソフトウェアのSCSIリザーブコマンド発行
- 112 実行結果の判定
- 113 ボリューム属性の確認
- 114 ベア再同期コマンド
- 115 ベアスワップ再同期コマンド
- 116 リザーブコマンドの失敗(Reservation Conflictステータス)
- 117 リザーブコマンドの成功(Goodステータス)
- 120 サスペンド状態(正常)
- 121 コピー状態
- 122 ベア状態
- 130 ベア再同期コマンドの正常終了
- 131 ベアスワップ再同期コマンドの正常終了

【図1】

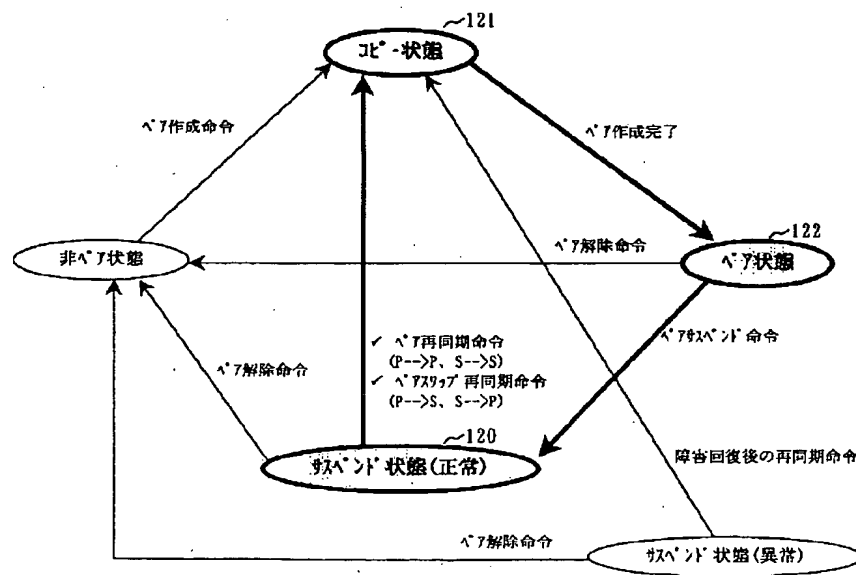
図 1



【図3】

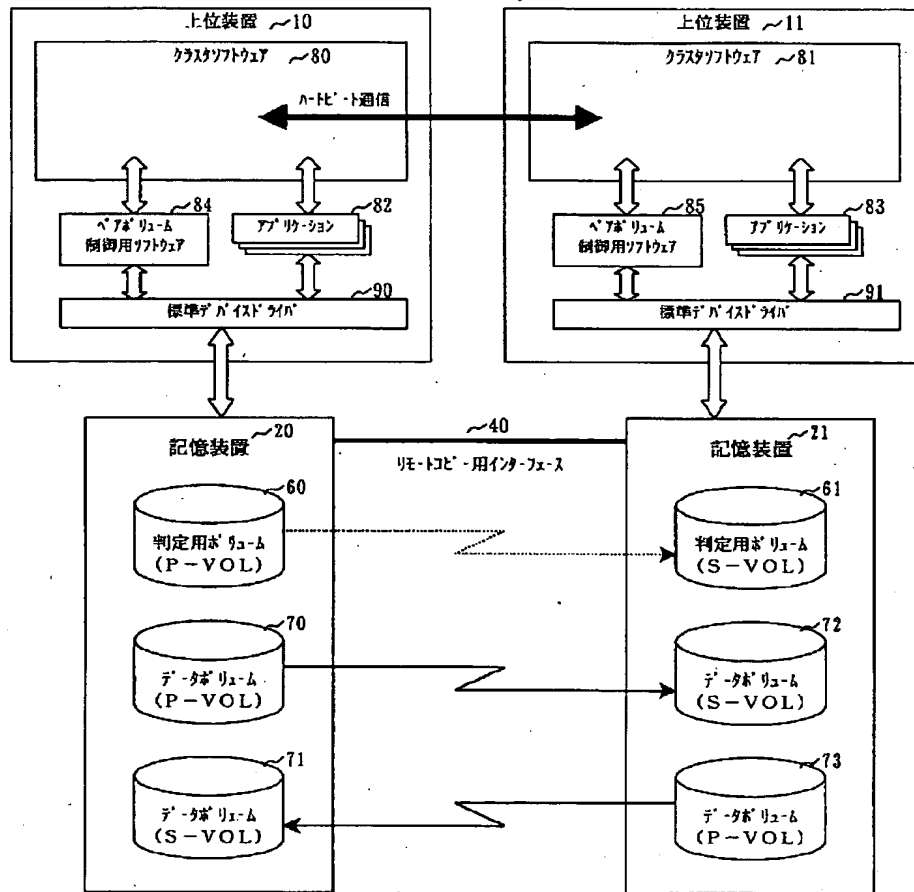
図 3

ペアボリュームの状態遷移図



【図2】

図 2



【図4】

図 4

ペア再同期コマンド発行時の記憶装置からのレスポンス

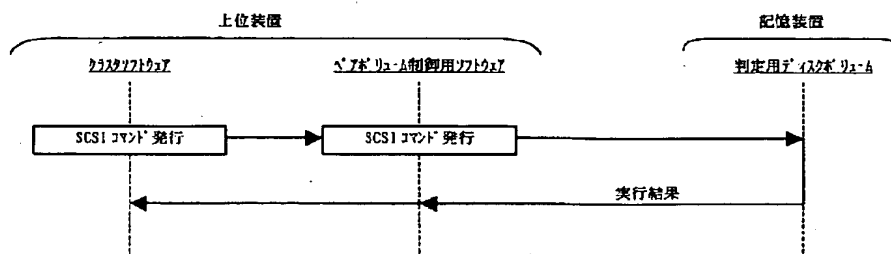
実行コマンド	自ボリユームの属性		記憶装置のレスポンス
ペア再同期	P-VOL	非ペアボリユーム	異常終了
		コピ-中	
		ペア状態	正常終了 ~130
		サスペンド状態(正常)	正常終了
	S-VOL	サスペンド状態(異常)	異常終了
		コピ-中	
		ペア状態	
		サスペンド状態(正常)	
ペアスワップ再同期	P-VOL	非ペアボリユーム	異常終了
		コピ-中	
		ペア状態	
		サスペンド状態(正常)	
	S-VOL	サスペンド状態(異常)	異常終了
		コピ-中	
		ペア状態	
		サスペンド状態(正常)	正常終了 ~131
		サスペンド状態(異常)	異常終了

【図6】

図 6

ペアボリユーム制御用ソフトウェアの動作

[リザーブコマンド以外]

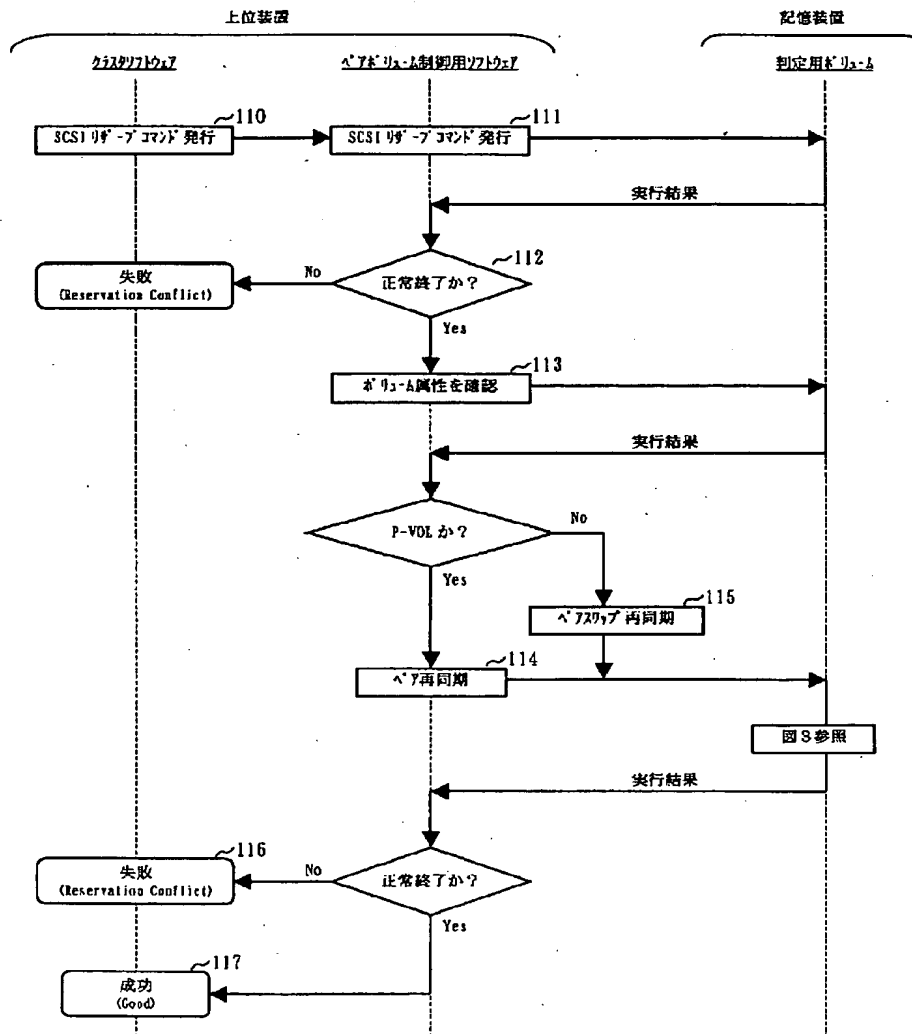


【図5】

図 5

ペアボリューム制御用ソフトウェアの動作

〔リザーブコマンド〕

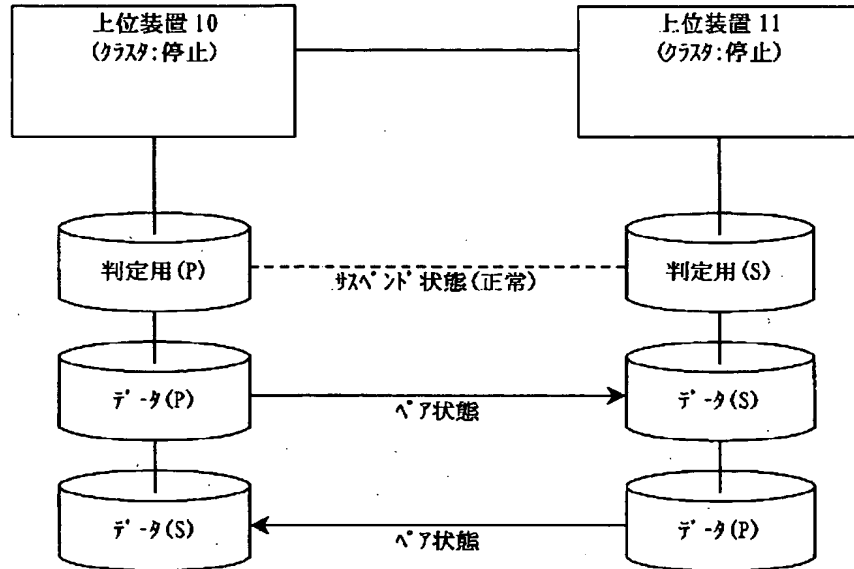


【図7】

図 7

初期状態

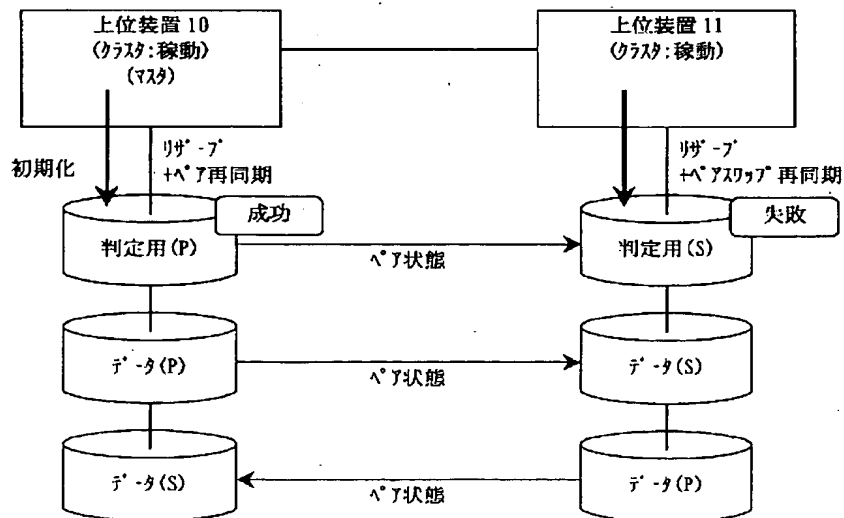
判定用ペアボリュームの属性：サスペンド（正常）



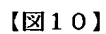
【図8】

図 8

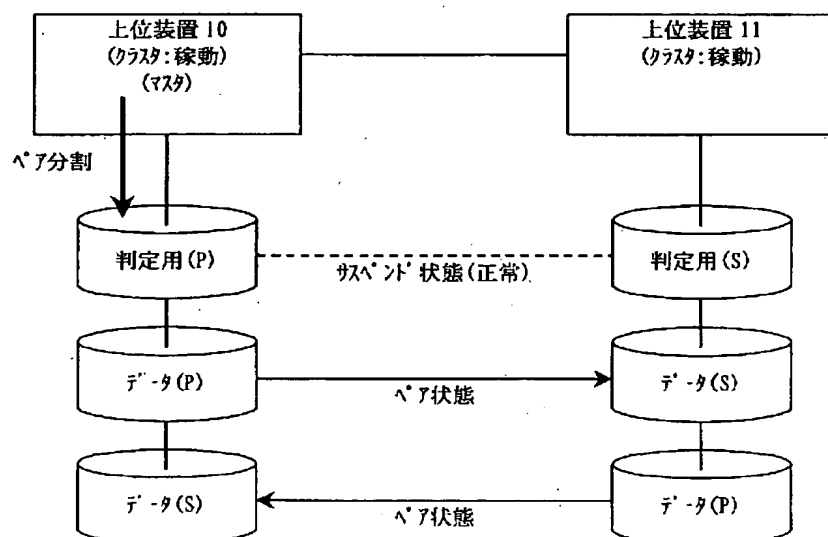
判定用ペアボリュームの属性：ペア状態



判定用ベアポリュームの属性：ベア状態



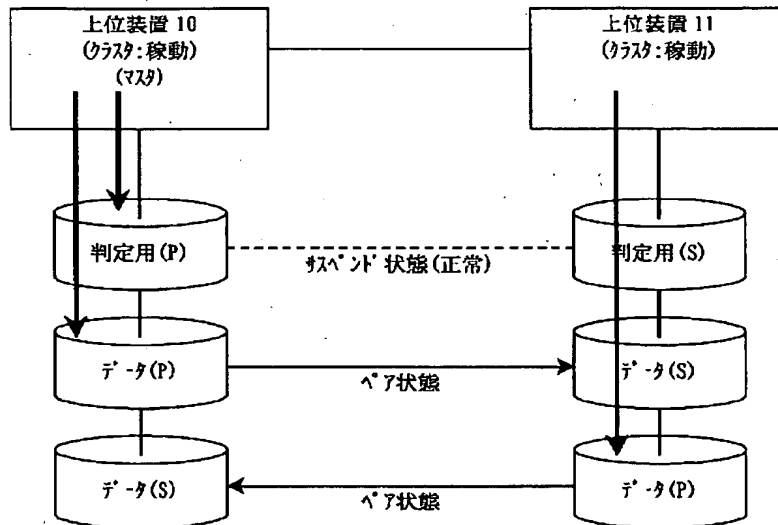
判定用ペアボリュームの属性：サスペンド（正常）



【図11】

図 1 1

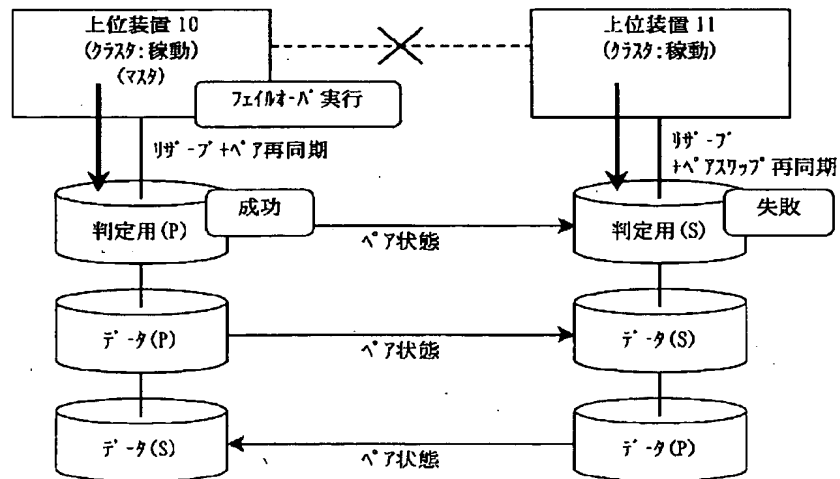
判定用ペアボリュームの属性：サスペンド（正常）



【図12】

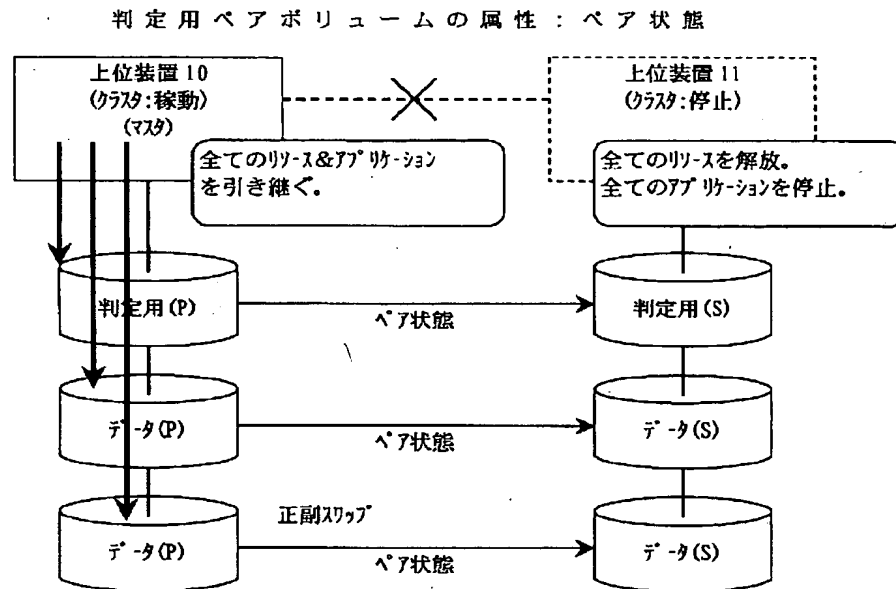
図 1 2

判定用ペアボリュームの属性：ペア状態



【図13】

図 1 3



フロントページの続き

(72)発明者 中野 俊夫
神奈川県小田原市中里322番地2号 株式
会社日立製作所RAIDシステム事業部内

Fターム(参考) 5B065 BA01 CA02 CC03 EA33 EK03
ZA01
5B082 DE07 FA17 GB02